Banff Challenge 2a Problems – Statistical Issues Relevant to Significance of Discovery Claims

W. Fisher, T. Junk, J. Linnemann, R. Lockhart, L. Lyons

November 23, 2010

Abstract

The workshop at the Banff International Research Station, 10w5068, on "Statistical Issues Relevant to Significance of Discovery Claims", raised several interesting issues that are best illustrated with concrete examples that participants can try out and discuss the issues that arise. This document provides instructions for trying out two such examples, which are meant to simulate the task of discovering new particle or phenomena in high-energy physics experiments.

1 Introduction

The Banff discovery challenge is designed to follow on to BIRS Workshop 10w5068, in which several interesting issues arose relating to discovery claims. Among these are the incorporation of uncertainty in the values of nuisance parameters, the utility of 3σ and 5σ significance requirements in HEP, computational difficulties raised by these stringent significance requirements, and tools for shortening the calculations. Also important is the the "Look-Elsewhere Effect", also known as the effect of multiple testing, or the "trials factor", which arises because the test hypothesis is not a simple hypothesis but includes extra parameters not present in the null hypothesis. We would like to put some of these ideas and questions to the test with examples that are realistic enough to capture the important parts of the statistical procedures for claiming evidence or discovery, while at the same time reducing the programming and computational requirements for full participation. We do realize that we would like to encourage participants to propose computationally efficient solutions to very challenging problems, but at the same time we would like to broaden the participant pool as much as possible, hence the structure of the problems and the requirements.

In a high-energy physics collider experiment, counter-rotating beams of particles are focused to a common collision area. Many different things can happen when particles collide. Usually several processes are involved in the collisions, most of which are known and well studied, and hence are not interesting, while

others are yet to be discovered (and which may or may not exist). Each collision which causes the detector electronics to decide to read out the detector is called an "event". Events may be caused by well studied and known processes, and are called "background" events. Events may also be caused by processes that have yet to be discovered. Physicists seek to collect samples of interesting events that can be used to convince others of the presence of new processes, so that they can claim discovery. Usually a physicist has some specific idea of a speculative, unconfirmed, new process that has a chance of being present in his data. Events that are caused by this process are called "signal" events. Usually physicists seek one kind of signal process at a time, although the true composition of the data events is unknown, and more than one kind of signal process may contribute to the data, in addition to those considered as backgrounds. We will keep the discovery challenge problems simple and only test the cases in which background processes and at most one signal process may be present.

We cast the problems as hypothesis tests. The null hypothesis (H_0) consists of the claim that the data consist only of background events. The test hypothesis (H_1) consists of the claim that not only are the background processes contributing, but also a signal component is present. The kind of signal sought is ideally specified before the data are collected and analyzed, and this will be the case here.

The detectors used in experimental particle physics are large and complex. They measure multiple properties of each event, such as the number of particles produced, their energies, their directions, and their particle type (energetic pions, electrons, photons, muons, protons, and many more exotic kinds of particles are produced regularly in these events). An example picture of an event is shown in Figure 1. In this picture, computer reconstructions of the paths left by the particles as they travel away from the collision in the center of the detector are shown. Energy deposited in the calorimeters isn't shown in the particular event display, but electrical signals in the outer muon detectors are. A great many properties are measured for each event.

These measured properties can be used to differentiate signal events (if they exist) from background events, usually very imperfectly. Particle physicists usually distill the measured properties of the events into one or two numbers (one in our case) that has an optimal separation of the distributions of signal and background events. Often neural networks, decision trees, and other kinds of dimension-reducing techniques are used to produce a single quantity as a function of the many measurements made on each event. The events are assumed to be Poisson distributed when counted, and this is true for any subset of the data that are collected based on fixed requirements used to select the subsets. The single number measured on each event is a mark in a marked Poisson process.

Background events are usually produced much more copiously than signal events. The differences in the distributions of the mark for the background component(s) and the proposed signal component, as well as *a priori* knowledge of the background rates (which may be poor), are used to make the claim of evidence for the signal component.

Two challenge problems are posed below. The main difference between the two problems is that the first one proposes models for the probability density of the marks for the signal and the background events that are simple, analytic functions. The second problem proposes models of signal and background distributions that are estimated using a Monte Carlo program. This latter case is more typical in experimental particle physics, as the predictions of the expected rates of signal and background events as functions of the mark involve integrals over positions, particle energies, sums over particle counts, and other quantities. These integrals are most easily evaluated with Monte Carlo programs which produce simulated events that are meant to mimic the data under H_0 and H_1. Particle physicists are often skeptical of whether their Monte Carlo programs are making unbiased predictions, and assign uncertainties to these predictions accordingly. The two hypotheses are compound hypotheses – they depend on the values of uncertain nuisance parameters.

2 Banff Discovery Challenge Problems

Both problems below provide specifications for H_0 and H_1 by defining the expected signal and background probability density functions of the marks and the expected prediction of the total numbers of events with uncertainties. Each problem's data also include a set of simulated data outcomes. These are lists of the marks for each event for each experimental outcome. A total of 20,000 simulated experimental outcomes (also called "datasets") is provided for each problem.

Some of the datasets will be generated from H_0, and others will be generated from H_1. Some of the datasets containing signals will be generated with multiples of the signal rate other than the standard ones specified below. In a real experiment, only one dataset is produced by the experimental apparatus, and it must be analyzed by itself with the help of the prior model information, which comes from subsidiary measurements and the work of theorists. **Challengees should not use any property of any simulated dataset to help interpret any other simulated dataset.** As the simulated datsets are drawn from different parent distributions, it would be risky anyhow to do this.

For the two problems described below, the challengee should provide the following

- For each of the problems, the power of the test for claiming evidence should be reported (physicists call this the "sensitivity"). The power is expressed as how often the challengee would claim evidence if a signal were truly present at the rate specified in the problem statement, using their technique adjusted such that the Type-I error rate is no more than 0.01.

- A description of the method used so that a document can be prepared describing each submission and its performance on the challenge problems. The description should be detailed enough so that the method can

be reproduced. Please describe any differences in the method used to address the two problems. If the extra credit intervals are computed, please provide a description for how these are obtained as well.

For the two problems described below, the challengee should do the following **for each simulated dataset**:

- A yes-no decision is to be made as to whether or not to claim that there is evidence in that dataset of the signal proposed in H_1. The desired Type-I error rate is 0.01. That is, no more than 1% of datasets provided that were generated using only H_0 should be flagged as having evidence for the signal process.
- One of the following should be provided:
  - The $\mathcal{P}$-value for H_0, *or*
  - The Bayes Factor, *or*
  - Some other quantity used to make the decision that can be described to the judges
- For problem 1, the challengees should compute for each simulated dataset for which they claim evidence, the location of the peak found (see the problem description below), and a 68% interval for the peak position of the signal found.
- *Extra Credit:* The challengees should compute 68% [confidence or credibility] intervals for the signal rate for each simulated dataset.

2.1 Problem 1: A Gaussian Signal Peak on an Exponential Background

For this problem, the simulated data samples are drawn from the following density functions[1]. The background density function is

$$B(x) = Ae^{-Cx} \tag{1}$$

where x is the mark of the event[2]. We will restrict the domain of x to be between 0 and 1. We choose the values $A = 10000 \pm 1000$ and $C = 10.0 \pm 0$. The background rate parameter A is drawn from a truncated Gaussian prior of width 1000, and truncated so that $A \geq 0$.

The signal density function is

$$S(x) = De^{-(x-E)^2/2\sigma^2} \tag{2}$$

[1] As these functions are not normalized to unit area, a better phrase is "intensity functions", but "intensity" means something else to a physicist.

[2] Physicists may want a concrete example – the dijet invariant mass of an event in a Higgs boson search is a typical choice for x, although neural network outputs are much more common these days.

We specify that $D \geq 0$, and that $\sigma = 0.03$, in our generation of simulated datasets. All signals in the simulated datasets have $0 < E < 1$. We call E the peak position, and we ask that challengees provide their best estimates of E for each dataset for which evidence is claimed, plus a 68% CL interval containing their best estimate. In computing intervals for D and E, you should take the above specifications as boundaries of the "physical region", that is, it is impossible to have a negative signal contribution, and the peak position range may be limited by another experiment's limit or perhaps a theoretical argument.

Because E can take any value in the range described above, the effect of multiple testing (the "Look Elsewhere Effect") is part of what we want the challengee to address in their decisions of whether to claim evidence for a signal or not.

For the power ("sensitivity") calculations, we would like challengees to specify the fraction of the time they estimate their technique would produce an evidence claim for a signal for three cases. These cases are given by $(D, E) = (1010.0, 0.1); (137.0, 0.5);$ and $(18.0, 0.9)$.

Each simulated dataset will be numbered. Please provide your decision results in an ASCII file with these columns:

- Dataset Number
- Decision to claim evidence for H_1 over H_0. 1 means evidence is claimed, 0 means no evidence.
- The P-value (or Bayes Factor, or your choice of variable)
- Best estimate of E if evidence is claimed, or zero if not claimed
- Low edge of the 68% CL region of E if evidence is claimed, or zero if not claimed
- Upper edge of the 68% CL region of E if evidence is claimed, or zero if not claimed
- Lower bound on D at 68% CL if the extra credit is attempted
- Upper bound on D at 68% CL if the extra credit is attempted

A series of simulated datasets is availale at the following location

`http://www-cdf.fnal.gov/~trj/bc2prob1.dat.gz`

Download it to your computer, and gunzip it to unpack the contents (the unzipped filesize should be about 194 MB). There are 20000 randomly chosen simulated datasets, some with simulated signals present, others without, with different values chosen for D and E. The first line in the data file has two integer numbers on it: the Dataset Number (from 0 to 19999) and the number of collision events $n_{\rm obs}$ for that dataset. The next $n_{\rm obs}$ lines in the file are the marks x for those events. After the marks are listed, the next simulated dataset begins.

2.2 Problem 2: A Monte-Carlo Parameterized Example

We pose a second exercise to simulate the challenges facing experimental particle physicists on a routine basis. In this case, we do not provide analytic functions for the distributions of the mark x measured on each event, but instead provide Monte Carlo parameterizations of of these distributions. Furthermore, we choose to model a situation in which the null hypothesis H_0 predicts that events are produced with two processes, which give different distributions of x and have different production rates, and the test hypothesis H_1 further posits the existence of a third process, the "signal" process, which has yet another distribution of x and a different production rate.

The total number of expected events is predicted from subsidiary measurements and/or theoretical predictions, and we summarize this knowledge in a set of priors. For the first background, "background 1", the expected number of events is 900 ± 90, where the shape of the prior is a truncated Gaussian of width 90 events, truncated so the predicted number of background 1 events is non-negative. For the second background, "background 2", the predicted rate is 100 ± 100 events, again where the shape of the prior is a truncated Gaussian, this time of width 100 events, truncated so the predicted number of background 2 events is non-negative. If this information came from a subsidiary Poisson measurement we would provide a more detailed likelihood function, but often there are larger sources of uncertainty in extrapolating the measurement in a subsidiary measurement to make a prediction in the selected event sample. Often, there are multiple estimates of a background such as number 2 above, which disagree at a large level such as this. Treat the uncertainties on backgrounds 1 and 2 to be uncorrelated.

For the two background sources and the signal, we have Monte Carlo simulations which provide predictions of the distributions of the mark x for the three processes. These may be found at

`http://www-cdf.fnal.gov/~trj/bc2p2bg1mc.dat.gz`
`http://www-cdf.fnal.gov/~trj/bc2p2bg2mc.dat.gz`
`http://www-cdf.fnal.gov/~trj/bc2p2sigmc.dat.gz`

and contain 5000 simulated events each. There is only one kind of signal possible, and only its total rate is uncertain (whereas Problem 1 had two separate parameters). Simulated datasets, some containing signals at various rates, others only drawn from H_0, are provided in the same format as in Problem 1, in the following file:

`http://www-cdf.fnal.gov/~trj/bc2prob2.dat.gz`

Please provide your answers to this problem in the same format as your answers to Problem 1, where the power is to be computed with an expected signal total rate of 75 events. Figure 4 shows an example simulated dataset, binned as a histogram together with the background models, with a signal present.

3 Turning in Your Results

Please send your files of decisions and intervals to the judges: `trj@fnal.gov` and `wfisher@fnal.gov`, no later than December 10 2010, in order to give the judges time to evaluate the submissions and prepare a document for PHYSTAT 2011, which starts on January 17, 2011. If you have questions regarding the problem statements and what is required, please don't hesitate to contact Tom Junk, `trj@fnal.gov` and Wade Fisher `wfisher@fnal.gov`.

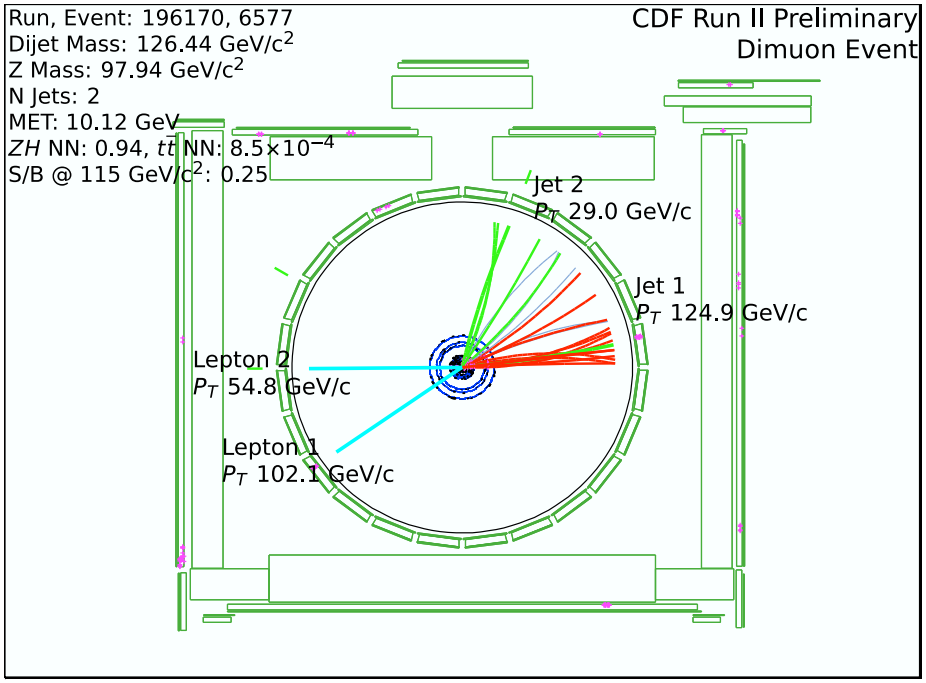

Figure 1: A display of an event collected by a high-energy collider experiment. The colliding beams travel perpendicular to the plane of the image, one into the page, one out of it. Tracks left by particles leaving the interacting region are shown in the inner portion of the picture, and electronic signals in the muon chambers (the outer straight-lined components) are also shown in pink. Energy deposits in the calorimeter are not shown.

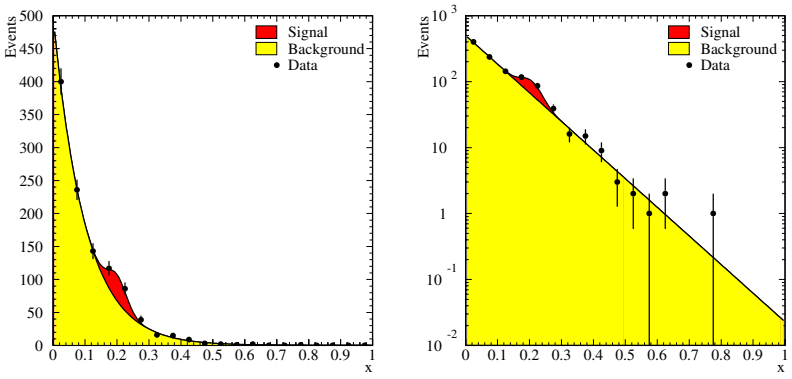

Figure 2: An example of an experimental outcome for Problem 1 which has been drawn from the test hypothesis H_1 with a signal rate that is large enough to be clearly visible above the background. The background used to generate the experimental outcome is shown as the light-shaded area, and the signal prediction is shown as the dark-shaded component stacked on top of the background. The simulated data are shown in bins of the mark x, as points with error bars. The size of the error bar in each bin is shown as the square root of the simulate data contents, as is the custom in particle physics plots (but not necessarily in analyses). The signal is chosen to be centered on $E = 0.2$. The left-hand plot has a linear vertical scale while the right-hand plot is the same on a logarithmic vertical scale.

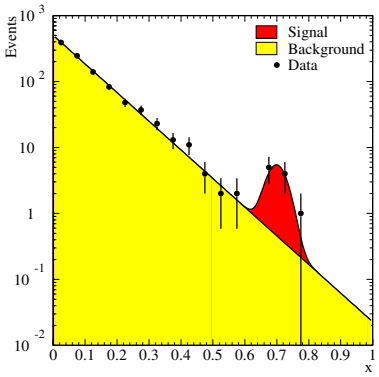

Figure 3: An example of a signal-like outcome for Problem 1 but for $E = 0.7$. See the caption of Figure 2 for an explanation of the items in the plot. Only the logarithmic scale plot is shown as the signal is not visible on a linear scale.

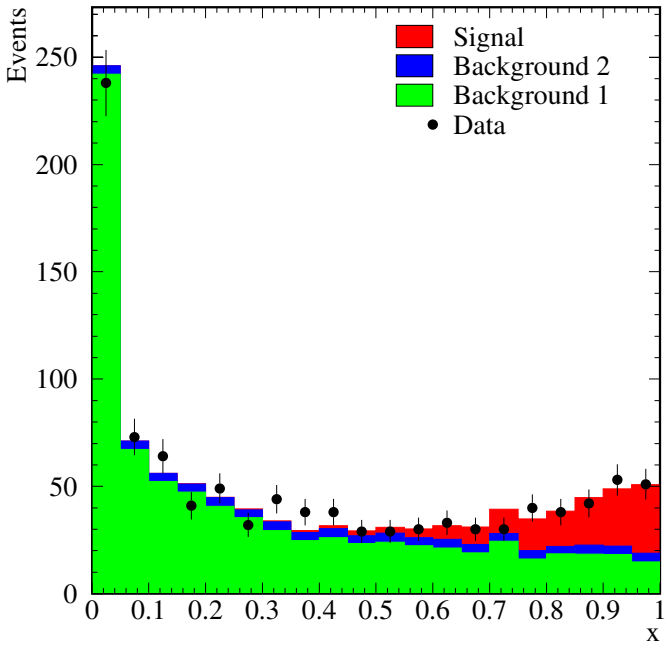

Figure 4: An example of a signal-like outcome for Problem 2. See the caption of Figure 2 for an explanation of the items in the plot. Two background sources are present, as well as the signal. The predictions in each bin are shown stacked.